Mellanox TECHNOLOGIES
www.mellanox.com

IBM Business Partner

IBM
www.ibm.com

CASE STUDY

Delivering Unmatched Messaging Performance with IBM LLM Software and Mellanox ConnectX® 10 Gigabit Ethernet Products

Commodity server and communication components are used to satisfy the extreme throughput and low latency needs of a stock exchange environment.

Situation

IBM's WebSphere MQ Low Latency Messaging (WMQ LLM) is a transport fabric product engineered for the rigorous latency and throughput requirements of today's financial trading environments. The transport provides one-to-one, one-to-many and many-to-many data exchange. It also exploits the IP multicast infrastructure to ensure scalable resource conservation and timely information distribution.

Designed to dramatically improve throughput and reduce latency while ensuring system reliability, WebSphere MQ Low Latency Messaging can help financial services organizations enhance the responsiveness of their existing trade infrastructure while developing new solutions for emerging business opportunities.

The Stock Exchange Reference Architecture (SXRA) application was developed in order to demonstrate the benefits of the WebSphere MQ Low Latency Messaging transport in a market data environment. This demonstration application shows how WebSphere MQ Low Latency Messaging can satisfy the challenging messaging requirements of a stock exchange in today's market data environment:

- Reliable message delivery – Reliable multicast messaging with fine-grained control of message delivery assurance, augmented by message persistence at wire speeds for message recovery and auditing.
- Extreme performance – A unique method of message-to-packet mapping enables delay-free, high-speed data delivery of hundreds of thousands (up to several millions) of messages per second, at microsecond latencies.
- High availability with fast failover – The Reliable and Consistent Message Streaming (RCMS) component provides highly available message streaming for fault tolerance. RCMS detects a component failure and migrates the data streaming from a failed to a backup application instance.
- Consistent message ordering – The total ordering feature of RCMS enforces a consistent order of message delivery from a number of independent data transmitters to multiple receivers, so all receivers see exactly the same order of incoming messages.
- Monitoring - WebSphere MQ Low Latency Messaging provides detailed visibility into the status of transmitters, receivers and latency.
- High Speed Interconnects – For the best possible latency and throughput, support for InfiniBand and 10 Gigabit Ethernet. The explosion of market data rates is exhausting the capacity of 1 GbE networks.

This case study discusses how the SXRA application uses WebSphere MQ Low Latency Messaging and Mellanox ConnectX-2 EN with RoE (RDMA over Ethernet) 10GbE controller satisfy the demanding reliability, throughput and latency needs of the stock exchange environment.

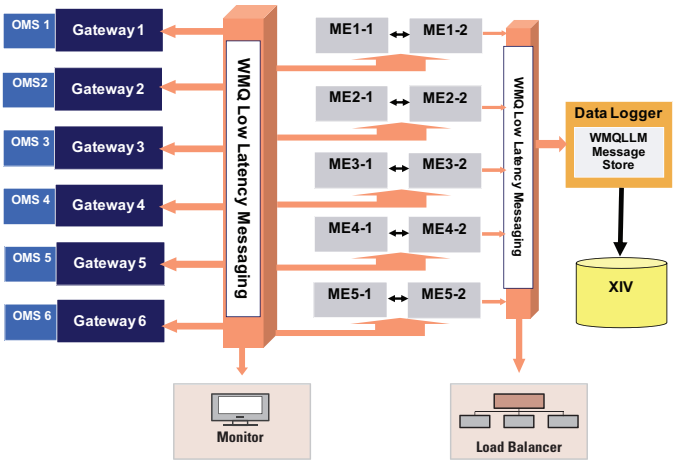
Solution Architecture

The SXRA application simulates the main components found in a stock exchange. The application consists of three components.

- The Ordering Management System (OMS) simulates a client that generates actions (Order, Cancel, CancelReplace, ...) and sends them to the exchange using the Financial Information eXchange (FIX) protocol. The OMS receives responses for the action it sent from the GW in FIX format. Since in most cases the OMS resides outside the exchange communicates with the GW is done over a simple TCP link and not RCMS.
- The Gateway (GW) receives actions from the OMS in FIX format and converts them to the internal format used by the exchange. The GW then processes the action and based on the symbol of the action forwards the action to the appropriate Matching Engine (ME). On the return path the GW receives responses (ActionResponse, Trade, etc.) from the ME. The responses are converted to FIX format and then sent to the OMS.

The Matching Engine (ME) receives actions from several GWs. Each action is processed according to its symbol using the appropriate order book. The outcome of the processing phase is an acknowledgment to the originating GW, as well as information on all the trades that the new action generated (if any). The trade information is sent to all the GWs which originated orders involved in that trade.

The basic architecture of the SXRA is presented in Figure 1 - SXRA Application. Gateways (six shown in the figure) receive orders from the OMS and send data to the Execution Venues (matching engines); in the figure there are 5 MEs where each ME is replicated twice. RCMS ensures that the input of messages to each ME is the same. The ME processes messages, which includes performing order matching and updating an in-memory order book. The output is then sent through RCMS back to the GWs which convert the data to FIX and forwards it to the OMS. RCMS automatically selects one of the tier members to act as primary and only this member will actually send the output to the destination.



Solution Implementation

The SXRA application has been implemented using the C/C++ APIs of the WebSphere MQ Low Latency Messaging product. A set of 6 Gateways and 5 highly-available ME pairs were run with messaging between these components using a Mellanox ConnectX EN with RoE -based 10GigE network. Mellanox ConnectX EN with RoE is based the RoCE (RDMA over Converged Ethernet) industry standard. ConnectX EN with RoE provides efficient RDMA transport to the SXRA application allowing message throughput and latency performance an order of magnitude better than other gigabit Ethernet and 10 gigabit Ethernet NIC solutions.

Benchmark tests have been conducted using hardware in IBM's Watson 590 lab with the following configuration:

- 8 IBM HS22 Blades
  - o 2 Intel® Quad Core Xeon® X5570 2.93 GHz 14 GB RAM
  - o Linux RHEL 5 update 3 (x86_64) 64-bit
- IBM H Type 3 Blade Center 8853
- 10 GbE with RoCE network
  - o BNT 10GbE switch Module for IBM BladeCenter
  - o Mellanox ConnectX-2 EN with RoE MT25428 10GbE adaptor
  - o OFED 1.5.1 with RoCE and ConnectX EN with RoE support

Results

The primary evaluation criteria for an exchange solution are the messaging performance and the high availability of the solution. Both of these criteria were evaluated in benchmark tests run using the SXRA application using the previously described configuration.

Messaging Performance

Tests were run with various message rates, tracking median, 50th, 90th and 99th percentile latency values. Latency was measured from the reception of a message from the OMS at a GW to the reception of a confirmation from an ME sent to the GW. This latency measurement includes the full processing of FIX messages in the GW. The results obtained are shown in Table 1 - SXRA latency results.

Rate (Orders/sec)	Average µs	50% µs	90% µs	99% µs
6,000	17	16	17	18
60,000	15	15	15	17
300,000	16	15	19	32
600,000	30	25	55	91
1,800,000	83	80	130	176
3,000,000	122	93	268	498

Table 1 - SXRA latency results

The full SXRA test was not run over standard 1 GbE networking. However, the higher message rates tested would not be possible on 1 GbE, and expected latency values would be an order of magnitude higher at all message rates. Also, performance results improve on those achieved using standard IP over 10 GbE, further demonstrating the value of using WebSphere MQ Low Latency Messaging with Mellanox ConnectX-2 EN with RoE adapters with OFED 1.5.1 and RoCE support.

High Availability

Failure recovery tests were performed to ensure that RCMS allowed the SXRA application to recover gracefully from any single failure or multiple independent failures. In addition the tests monitored the estimated time it took to recover from a failure.

The tests included the following steps

1. Kill one (or more) of the running ME processes.
2. Observe the latency introduced by the failure.
3. Restart the terminated ME processes.
4. Wait for RCMS to synchronize the new processes.
5. Observe the latency introduced by the synchronization process.
6. Go back to 1.

To verify that failure recovery completed successfully a monitoring mechanism has been added to the SXRA application to alert whenever any data (Order, Acknowledge, or Trade) has been detected to be missing.

Test results demonstrated that the failure process completed successfully without any message loss. The recovery times varied depending on the rate of orders being generated and sent through the system, with higher message rates naturally resulting in longer recovery times. Failure detection was typically on the order of 50 ms, with recovery times remaining well below 1 second.

Conclusions

The test results demonstrate that the WebSphere MQ Low Latency Messaging product running over low latency Ethernet can be used to implement the high performance, highly available messaging infrastructure needed for the next generation implementations of exchange systems. The RCMS component provides the facilities needed to implement highly available message streams in the applications. Support for next-generation communication fabrics allows applications to achieve the lowest possible latencies at high message volumes required to meet the trading targets that their participants require. This is demonstrated in this use case using the Mellanox ConnectX-2 EN with RoE adapters with OFED 1.5.1 and RoCE support, permitting efficient RDMA communications over a 10 GbE network.